

Development of improved K-means clustering for health insurance claims

¹Stephen G. Fashoto, ²Adekunle Adekoya, ³Jacob A. Gbadeyan and ⁴J. S. Sadiku

¹Kampala International University, Kampala Uganda

²Redeemer's University, Ede Osun State Nigeria

³University of Ilorin, Ilorin, Nigeria

Healthcare insurance, delivered via the National Health Insurance Scheme (NHIS) is a veritable tool for making quality healthcare available to the majority of Nigerian citizens, irrespective of income status. This scheme, however, faces imminent collapse if an effective way of grouping health insurance claims is not found. Health insurance claims account for a significant portion of all claims received by insurers amounting to billions of naira annually. Thus, this study focused on application of data mining techniques that could help to drastically reduce the time spent on segmenting health insurance claims in health insurance administration in Nigeria. The specific objectives were to: i) develop and implement “real-time assignment” K-means clustering algorithm on health insurance claims and ii) develop and implement “modified initialize scheme” K-means clustering algorithm on health insurance claims.

The proposed algorithms for the improved K-Means Clustering are implemented using JAVA.

The improved K-means clustering algorithm was employed in solving the segmenting problems of the different health insurance claims.

This study concluded that the proposed algorithms are superior to the traditional K-means clustering in terms of convergence and accuracy. The resulting clusters will be used in further studies for developing the supervised classification approach.

KEYWORDS: *K-means clustering, Distance functions, health insurance, health insurance claims.*

1. Introduction

Government in most countries lay emphasizes on healthcare service delivery as their basic function. The assignment is overwhelmed with complications especially in Nigeria. One of the ways to conquer the challenges of the health insurance sector in Nigeria is by applying software techniques that will help to minimize errors in the health insurance claims [1]. The main focus of this study is how to use data mining techniques to drastically reduce the time spent and inaccuracy in health insurance administration in Nigeria for grouping health insurance claims. Health Insurance is any approach that makes it possible for people to receive healthcare services or products without the need to pay for such services and products at the point of care [2].

The health insurance industry has historically been a growing industry. The role of the industry in the economic and social life of any country is so crucial that it cannot be overemphasized. But ever since its beginning as a commercial enterprise, the health insurance industry has been facing difficulties with segmenting health insurance claims. Insurance fraud is very costly and has become a worldwide source of concern in recent years. A sizeable part of insurance claims are attributed to be Fraudulent, whose value amount to billions of Naira yearly. Nowadays, great efforts are being made to develop models to identify potentially fraudulent claims for special investigations using data mining technology [3].

Health insurance is a product offered to a person, a family member, or employees of an organization subscribing for some healthcare covering for a fee. It is usually an arrangement designed to assist subscribers to minimize payment for healthcare service delivery [3].

[4] deciding the number of groups of patients is imperative in healthcare industry and it requires high accuracy. There are several ways to determine the optimal value of K clusters but the most popular one is running the algorithm with different values of K. But this study also addressed two of the limitations of K-means, which is improving the accuracy and convergence rate.

One of the main challenges of the health insurance sector especially in Nigeria is on how to design a robust system that will be able to group health insurance claims without having to rely solely on the internal control or auditing system to reveal the characteristics of the claims.

By tradition, most companies rely on its internal control system and internal auditing system for grouping claims. Once the internal control and the internal audit system fail then different information technologies techniques are tried to safeguard further occurrence, but it is rather unfortunate that most of these solutions are not sufficient.

Data mining techniques, on the other hand, are holding out a great promise as regards their ability to improve accuracy of grouping health insurance claims. Data mining combines powerful analytical techniques with knowledge to turn the data already acquired into the information and insight needed to identify probable instances of the health insurance claims [5].

The optimal value of a cluster is referred to as K and selecting the appropriate value of K is a difficult task without a prior knowledge of the input data. K can be determined by performing clustering for a range value of K and after that the least value of K can be selected for cluster validity measure. This procedure is computationally intensive when the actual number of clusters is large [6].

A number of researchers have attempted to improve the traditional K-means clustering shows in table 1 which could be applied in a number of real-life classification scenarios. The health insurance claims and Iris dataset is such a domain where K-means clustering has proved to be useful. A few authors have attempted to utilize improve version of the K-means Clustering.

Table 1. Related literature on Clustering Technique

Author	Application domain	Clustering technique	Dataset
[7]	Healthcare insurance fraud	K-means	Life claims payment data
[8]	Healthcare insurance fraud	Resolution based clustering	Policy holder attributes data
[9]	Healthcare fraud	K-means	Healthcare payments data

This may perhaps be as a result of cluster validation which is an active research focus that stresses two essential issues which must be tackled and they are: how to approximate the number of clusters in a data set and evaluation of clustering algorithms [10]. Data mining appears to be an efficient method in supervising transaction [11]. Sadly K-means is very sensitive to centriods. If the initial partitions are not chosen carefully, the computation will run the chance of converging to a local minimum rather than the global minimum solution. Because of initial starting points generated randomly, K-means does not guarantee the unique clustering results [12].

So far, in Data Mining research there seems to be no general approach in existence on how to normalize a dataset and so the choice is based on the discretion of the user [13]. The erroneous choice of picking or determining an optimal value of K for a definite dataset in K-means clustering algorithm will produce a wrong decision for the partitioning scheme. The difficulties of choosing

the number of clusters that best fits a dataset as well as assessing the clustering results has been an issue in many studies.

Several new clustering algorithms have been proposed to overcome the drawbacks of K-means clustering

- (i) The genetic weighted K-means algorithm which is a hybridization of genetic and weighted K-means algorithm. These perform better than the K-means in terms of the cluster quality and the clustering sensitivity to initial partitions [14]. These partially solve the problems of clusters with spherical-shape.
- (ii) According to [15] proposed a modified version of a well known traditional K-means clustering algorithm. The modified algorithm maintains all important characteristic features of the basic K-means and at the same time was able to partially eliminates generation of empty clusters.
- (iii) [12] proposed an approach to initialize cluster centres based on values for each attribute of the dataset. These approach is time costly and may not be applicable by keeping the K-means inherently simple structure.
- (iv) Use of one additional centroid approach is also proposed to cluster data objects. Several partitioning and merging process are used. Merging decision depends on the average mean distance where average distance between each cluster mean and each data object is determined. Since the least and closest clusters in average mean distance are merged in one cluster, this process continues until we get the final required clusters in an accurate and efficient way [16].
- (v) [17] proposed an efficient method for assigning data-points to clusters. The original K-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose- one similar to the K-means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original k-means algorithm. Hence there is no guarantee for the accuracy of the final clusters.

The "improved K-means clustering algorithm" was employed to discover the grouping of the different health insurance claims (cluster 1 and cluster 2) and iris dataset (iris setosa, iris versicolor and iris virginica).

The K-means Clustering on data mining is presented in section 2, Section 3 describes the algorithms of the proposed improved K-means clustering used in identifying the extent of segmenting claims in health insurance industry and Iris dataset for benchmarking purpose. The software requirements are presented in section 4. The results of the generic dataset and empirical datasets used are also presented and discussed in section 5. In section 6, the conclusion and the recommendations are presented.

2. K-means Clustering Algorithm

The K-means algorithm is extensively used to generate clustering of data because of its speed and simplicity [18]. It is a simple iterative method to partition a given dataset into a user-specified number of clusters, k [19]. The idea is based around clustering items using centroids. These are points in the metric space that define the clusters. Each centroid defines a single cluster, and each point from the data is associated with the cluster defined by its closest centroid. Then the algorithm iterates between two steps till convergence. The first step is data assignment. Here, each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data. The second step is relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions. The K-means algorithm is simple, easily understandable and reasonably scalable, and can be easily

modified to deal with streaming data. However, one of its drawbacks is the requirement for the number of clusters, K , to be specified before the algorithm is applied [19].

[20] presented an experimental study on K-means, SA (Simulated Annealing) and GA (Genetic Algorithm) and concluded that the SA and GA are better than K-means in terms of quality solution but the K-means is better than the SA and GA in terms of execution time when applied on segmenting of small datasets. The K-means algorithm and Artificial Neural Network (ANN) are incomparable when applied on large datasets.

2.1 Distance Functions

Distance functions in K-means clustering technique plays an important role. Different distance functions are provided to measure the distance between data objects. These two distance functions are as follows

2.1.1 Euclidean Distance Function

Euclidean distance is ordinary distance between two points that one would measure with a ruler. It is the most commonly used distance function [21]. This distance is given by Pythagorean formula. The Euclidean distance between the points a and b is the length of the line segment connecting them (a, b). In the Euclidean plane, if $a = (a_1, a_2)$ and $b = (b_1, b_2)$ then the distance is given by: $D(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$. Weakness of the basic Euclidean distance function is that if one of the input attributes has a relatively large range, then it can overpower the other attributes [22].

2.1.2 Manhattan Distance Function

In Manhattan distance function the distance between two points is the sum of the absolute differences of their coordinates. The Manhattan distance, D_1 between two vectors a, b in an n -dimensional real vector space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axis [21]. More formally, $D_1(a, b) = \|a - b\|_1 = \sum_{i=1}^n |a_i - b_i|$, where $a = (a_1, a_2 \dots a_n)$ and $b = (b_1, b_2 \dots b_n)$ are vectors.

In this study we use Euclidean distance metric instead of the Manhattan distance metric for the numeric attributes because of its popularity.

The reasons behind the popularity of the K-means algorithm are:

1. Its time complexity is $O(mkl)$, where m is the number of instances; k is the number of clusters; and l is the number of iterations taken by the algorithms to converge. Typically, k and l are fixed in advance and so the algorithm has linear time complexity in the size of the data set.
2. Its space complexity is $O(k+m)$. It requires additional space to store the data matrix. It is possible to store the data matrix in a secondary memory and access each pattern based on need. However, this scheme requires a huge access time because of the iterative nature of the algorithm. As a consequence, processing time increases enormously.
3. It is order independent. For a given initial seed set of cluster centers, it generates the same partition of the data irrespective of the order in which the pattern are presented to the algorithm.

Other reasons for the algorithm's popularity are its ease of interpretation, simplicity of implementation, speed of convergence and adaptability to sparse data.

3. Proposed Algorithms for Improved K-means Clustering

3.1 Comparison of pseudo-code for Batch Assignment K-means(BAK) and Real-time Assignment K-means(RAK)

```
void main()
begin
```

```

// inputs to this algorithm are :
// 1) set of clusters  $s = \{s_1, s_2, \dots, s_n\}$ 
// 2) set of observations or inputs  $x = \{x_1, x_2, \dots, x_n\}$ 
// Output from this algorithm :
// 1) Observations assigned to appropriate clusters
// 2) Performance statistics such as mis-classification rates
k_means(s,x) // function invokes here
end ;
func k_means(s,x)
begin
// s => set of clusters
// x => set of inputs or observations
initialize_kmeans(s,x)
t := 0 ;
EOC := false // a measure of convergence
NOT_EOC := false
while NOT EOC
for i = 1 to s.length()
for p = 1 to x.length()
is_a_member_of(x,p,s,i, s.length() , NOT_EOC )
end // for
end // for
if (NOT_EOC = true ) then
t := t + 1
else
EOC := true
end if
end while
end // end of k_means
BAK
func is_a_member_of(x,p,s,i, k , NOT_EOC ){
begin
D := sqrt((x[p]-m[i])^2) // abstract operation ...
for j:=1 to k

d := sqrt((x[p]-m[j]) ^ 2)
if(D > d) return false ;
next j
add_to_this_cluster(s,i,x,p,NOT_EOC ); // add this to a queue
return true ;
end ;

```

```

RAK
func is_a_member_of(x,p,s,i, k , NOT_EOC ){
begin
D := sqrt((x[p]-m[i])^2) // abstract operation ...
for j:=1 to k

d := sqrt((x[p]-m[j]) ^ 2)
if(D > d) return false ;
next j

```

```

if( i <> p ) then
    add_to_this_cluster(s,i,x,p,NOT_EOC ); //
    m := compute_mean_of_clusters(s)
end if
return true ;
end ;

```

3.2 Comparison of pseudo-code for Traditional Initialize K-means(TIK) and Modified Initialize K-means(MIK)

TIK

```

func initialize_kmeans(s,x)
begin
    randomly_assign_inputs_to_clusters(s,x)
    m = compute_mean_of_clusters(s)

end

```

MIK

```

func initialize_kmeans()
begin
    frequency_table otable = null; // new frequency_table();
    for (int i = 0; i < this.iNoOfInputs; i++) {
        int iClusterIndex = this.obuffer.get_ith_cluster_index(i+1);
        if (isEmptyCluster(iClusterIndex)) {
            oInputs[i].assign_node_to_cluster(iClusterIndex);
        } else {
            assign_to_the_closest_cluster(oInputs[i], otable);
        }
        Compute_mean_of_cluster()
    }
end

```

The proposed algorithms for the improved K-Means Clustering are implemented using JAVA.

3.3 Description of the Data Collected

The pertinent data to carry out the study is collected from HMO claims database and Iris dataset from University of California Irvine (UCI) repository. Though each of the tables has attributes in the original dataset, the table 2 and 3 show selected attributes of each table.

Table 2. HMO Claims Details

Attributes Name	Data type
Enrollee id	a unique identification number
Diagnosis	Nominal
HospitalName	Nominal
Class of treatment	Nominal
Amount Billed	Numeric
Amount Approved	Numeric

Year	Numeric
Month	Nominal

Table 3. Iris Dataset

Attributes names	Data type
Sepallength	Numeric
Sepalwidth	Numeric
Petallength	Numeric
Petalwidth	Numeric

4. Implementation Tools

4.1 Java Programming Language

Java programming language was used for the development of the three different K-means clustering algorithm (traditional, real-time assignment and modified-initialize) to determine the Mean Square Errors and Average Misclassification percentage errors via the Java Development kit (JDK), which is the compiler for Java programming language. Netbeans 6.8 IDE was used as the programming platform.

The major reason Java programming language was used is because of the researcher understanding of the language and its numerous capabilities.

4.2 Clustering Experiment

The clustering task of segmenting Iris datasets and health insurance claims datasets was done using the Java code for the Real-time Assignment K-means clustering (RAK), Traditional K-means clustering (TKM) and Modified-Initialize K-means clustering (MIK). Accordingly, the HMO experts have been consulted in setting the optimal value. They have suggested that the K value to be 2 (representing cluster 1 and cluster 2 health insurance claims) but k=3 were chosen for the Iris datasets. This cluster approach is experimented and evaluated against its performance in creating dissimilar clusters/segments when the default parameters are changed. According to the works of [23], the notion of “good” clustering is strictly related to the application domain and its specific requirements. Nevertheless the generally accepted criteria for validating the clustering results in different domains are the measures of separation among the clusters and cohesion within clusters (that is inter and intra cluster similarities respectively). So, for validating the clustering result of this study the Mean Square Error (MSE) and Average Misclassification Percentage Error (AMPE) were used. The experiment was performed for 100 runs with k=3 for the iris dataset and k=2 for the health insurance claims dataset.

5. Results and Discussions

5.1 Experimental Results on improved K-means Clustering

This sub-section provides a comparison of the traditional K-means clustering (Batch Assignment K-means (BAK) and Traditional Initialize K-means (TIK)) and the proposed algorithms (Real-time Assignment K-means (RAK) and Modified-Initialize K-means (MIK)) in terms of convergence and accuracy when clusters are of same size and same dimensions (Table 4 and 5). We use K-means as a guide to find the optimal solution to assign data objects to the correct cluster. BAK on the traditional K-means make use of the same centroid per iterations for the entire input one after the other while RAK make use of different centroids for different inputs per iterations, but if the inputs do not change cluster membership in iteration then RAK will behave like BAK.

RAK on the improved K-means outperform the BAK on the traditional K-means (Table 6 and 7).

MIK randomly select a cluster if the cluster is empty, assign the input to the cluster else the cluster is not empty then get the list of all such non-empty clusters then assign the input to the closest clusters that is among the list of all non-empty clusters while the Traditional Initialization K-means (TIK) randomly pick a cluster then assign an input to that cluster. And the MIK also outperform the TIK (Table 6 and 7). RAK makes the assignment to converge faster while MIK makes the initialization scheme to be more accurate.

RAK, TKM and MIK are applied to iris datasets and health insurance claims datasets in order to be able to analyze the accuracy of the proposed algorithms.

Iris Datasets

In this experiment, four dimensional dataset with Real-time assignment function were used in Java to generate the randomly assigned inputs to clusters on the iris datasets and Batch assignment function in Java to generate the randomly assigned inputs to clusters on the iris datasets. The iris dataset are made up of 150 data objects that are categorized into three groups (iris setosa, iris versicolor and iris virginica) and each data object has four numeric attributes(sepal length, sepal width, petal length, petal width) and each group has 50 data objects (Table 4).

Table 4. Description of Iris datasets

Datasets	Data sizes	Dimension	No. of clusters
Iris datasets for Real-time Assignment	150	4	3
Iris datasets for Batch Assignment	150	4	3

Health Insurance Claims Datasets

In this experiment, two dimensional dataset with Real-time assignment function were used in Java to generate the randomly assigned inputs to clusters on the health insurance claims datasets and Batch assignment function in Java to generate the randomly assigned inputs to clusters on the health insurance claims datasets. The health insurance claims dataset consists of 2,477 data objects that are categorized into two groups (cluster 1 and cluster 2) and each data object has eight attributes(Enrollee_id, hospital_name, year, month, class_of_treatment, Diagnosis, Amount_Approved and Amount_Bill) (Table 5).

Table 5. Description of Health insurance claims datasets

Datasets	Data sizes	Dimension	No. of clusters
Iris datasets for Real-time Assignment	2,477	8	2
Iris datasets for Batch Assignment	2,477	8	2

Average Misclassification Percentage Error (AMPE) and Mean Square Error (MSE) were used to validate the results of our proposed algorithms by computing RAK, MIK and TKM for iris datasets and health insurance claims datasets after running them for 100 runs, the average results were taken. The results of the comparison are given in Table 6. The error criteria such as mean square error (MSE), and average misclassification percentage error (AMPE) are assessed on the proposed improved K-means Clustering algorithm performance.

The formulas of error criteria are as follows: $MSE = \sum_{i=1}^N \frac{(y' - y)^2}{N}$.

Table 6. Misclassification Percentage Error comparison

Datasets	Algorithm	AMPE(%)
Iris	RAK	33.3716
Iris	TKM	35.4107
Iris	MIK	33.5250
Health insurance claims	RAK	31.08597
Health insurance claims	TKM	31.08599

Health insurance claims	MIK	31.08598
-------------------------	-----	----------

Table 6 shows the Average Misclassification Percentage Error comparison on the proposed algorithms (RAK and MIK) and traditional K-means.

Table 7 shows the Mean Square Error comparison on the proposed algorithms (RAK and MIK) and traditional K-means.

Table 7. Mean Square Error comparison

Datasets	Algorithm	MSE(%)
Iris	RAK	33.4716
Iris	TKM	35.5107
Iris	MIK	33.6250
Health insurance claims	RAK	38.5456
Health insurance claims	TKM	38.5458
Health insurance claims	MIK	38.5457

The iris dataset and the health insurance claims dataset AMPE and the MSE for 100runs are shown in Table 6 and 7 respectively. The result on iris dataset on table 6 reveals that there are 33.3716% AMPE using RAK, 35.4107 for MIK and 33.5250% for TKM. 31.08597% AMPE for RAK, 31.08598% for MIK and 31.08599% AMPE for TKM on health insurance claims dataset. Thus the classification accuracy on iris dataset is 89.33% for RAK and 88.67% for MIK and TKM while the classification accuracy on health insurance claims dataset is 68.1403% for RAK, 68.1402% for MIK and 68.1401% for TKM.

The results on iris dataset on Table 7 reveals that there are 33.4716% MSE using RAK, 33.6250% using MIK and 35.5107% using TKM while the MSE from health insurance claims dataset are 38.5456% for RAK, 38.5457% for MIK and 38.5457% for TKM. From Table 6 and Table 7 the RAK outperform the MIK and TKM in terms of AMPE and MSE.

6. Conclusion

This study was conducted using improved K-means clustering. The initial data collected from the Health Maintenance Organization (HMO) and the University of California Irvine (UCI) repository did not incorporate the target class for this study. The clustering module was conducted using the “real-time and modified initialize K-means” clustering algorithm. This can be used for segmenting the data into the target classes of Health insurance claims dataset from the HMO and iris datasets from the UCI repository. Applications on well-known benchmark dataset (iris dataset) and empirical datasets (Health insurance claims dataset) correlating it with traditional K-means algorithms, it was found that the results obtained yielded a better convergence and accurate results.

This study established an improved Real-time assignment K-means clustering approach to solve the problems in National Health Insurance Scheme that can be used in categorizing health insurance claims.

And for further studies, the resulting clusters can be used by supervised learning classification techniques such as Hidden Markov Model (HMM), Multilayer Perceptron (MLP) and Self-Organizing Mapping (SOM) for further investigation.

References

- [1] Fashoto, S.G., Gbadeyan, J.A. and Sadiku, J.S. Application of Data Mining technique to fraud detection in Health Insurance Scheme using Multilayer Perceptron. In the Proceedings of the 2nd annual International Conference on E-leadership of IEEE in University of Pretoria, Pretoria, South Africa, 2013a.
- [2] Care Net Nigeria, *Health Insurance Report and Health Insurance Affairs* January 2002 to October 2010 issues. www.carenet.info/Resources (accessed on 21/01/11), 2011.
- [3] Fashoto, S.G., Owolabi, O., Sadiku, J.S., and Gbadeyan, J.A. Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm. *Australian Journal of Basic and Applied Sciences*, 2013b, 7(8): 140-144.
- [4] Nguyen, D.T., Nguyen, G.T. and Lam, V.T. An Approach to Data Mining in Healthcare: Improved K-means Algorithm. *Journal of Industrial and Intelligent Information*, 2013, Vol. 1, No. 1.
- [5] Yang, W.S. and Hwang, S.Y. A process-mining framework for the detection of healthcare fraud and abuse. *Expert systems with Applications*. 2005.
- [6] Du, K.L. Clustering: A Neural Network Approach. *Journal of Neural Networks*, 2010, 23(1), pp. 89-107.
- [7] Thiprungsri, S. and Vasarhelyi, M. Cluster Analysis for Anomaly detection in Accounting Data An Audit Approach. *The International Journal of Digital Accounting Research*, 2011, volume 11.
- [8] Xiaoyun, W. and Danyue, L. Hybrid Outlier mining algorithm based evaluation of client moral risk in insurance company. The second IEEE International Conference on Information Management and Engineering (ICIME), 2010, pp. 585-589.
- [9] Little, B., Rejesus, R., Schucking, M. and Harris, R. Benford's Law, Data Mining and financial fraud: A case study in New York State Medicaid data, 2008, Volume 9, pp. 95—204.
- [10] Peng, X., Zhang, L. and Yi, Z. Constructing l2-graph for subspace learning and Segmentation, 2012.
- [11] Sabau, A.S. Survey of Clustering based Financial Fraud Detection Research. *Informatica Economica*, 2012, 16(1), pp. 110 – 122.
- [12] Khan, S.S. and Ahmed, A. Cluster center initialization for K-means algorithm. *Pattern Recognition Letters*, 2004, 25(11), pp. 1293-1302.
- [13] Karthikeyani, V.N. and Thangavel, K. Distributed Data Clustering: A Comparative Analysis. *Foundations of Computational Intelligence* 2009, 6, pp. 371-397.
- [14] Wu, X., Kumar, V., Quinlan, R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, J., Ng, A., Liu, B., Yu, S., Zhou, Z., Steinbach, M., Hand, J. and Steinberg, D. *Top 10 Algorithms in Data Mining*. Springer-Verlag, London, 2008.
- [15] Pakhira, M. A Modified K-means Algorithm to avoid empty clusters". *International Journal of Recent trends in Engineering*, 2009, 1(1), pp. 201-206.
- [16] Alnaji, P. Refining initial points for K-means clustering. *Proceedings of Fifteenth International Conference on Machine Learning*, San Francisco, CA, Morgan Kaufmann, 2011, pp. 91-99.
- [17] Fahim, A., Salem, A.M., Torkey, A. and Ramadan, M.A. An Efficient enhanced K-means clustering algorithm. *Journal of Zhejiang University*, 2006, 10(7): pp. 1626–1633.
- [18] Graham, M.S. Neural network-based systems for handprint OCR applications. *Image Processing*, 2006, vol. 3, no. 8.
- [19] Pham, D.T., Dimov, S.S. and Nguyen, C.D. Selection of K in K-means Clustering. *Journal of Mechanical Engineering Science*, 2005, 219, 103-119.
- [20] Al-sultan, K. A Tabu search approach to the clustering problem. *Pattern Recognition*, 1995, 28(9), pp. 1443-1451.
- [21] Antoni, M. The case for approximate Distance Transforms" Presented at SIRC 2002 – The 14th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand December 3-5th, 2002.
- [22] Randall, W.D. and Martinez, T.R. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research* 1997, 6, 1-34 Submitted 5/96; published 1/97, AI Access Foundation and Morgan Kaufmann Publishers.
- [23] Halkidi, M. and Vazirgiamis, M. Evaluating the validity of clustering results based on density criteria and multirepresentatives, 2001.